基于 DSpace 构建机构仓储的备份与恢复

陈 和
（厦门大学图书馆 厦门 361005）

［摘要］针对导致基于 DSpace 构建的机构仓储系统出现故障或错误的各种可能原因，以及仓储系统的数据存储特点，分别介绍和分析了三种类型的数据备份与恢复方法，即使用导入与导出工具进行备份与恢复、数据存储级备份与恢复、文件系统级备份与恢复。并在此基础上，推荐使用一种简单实用的自动异地文件系统级备份方法。
［关键词］DSpace 机构仓储 备份 恢复
［分类号］TP39

The backup and recover of Institutional Repository Base on DSpace

Chen He
(The Xiamen University Library, Xiamen, 361005, China)

[Abstract] According to different possible factors which induce error of institutional repository based on DSpace system, and to the data storage specialty of institutional repository, three type methods of data backup and recover are introduced and analyzed, which are backup and recover with importing and exporting tools, backup and recover based on data storage, backup and recover based on file system. Finally, a simple applied method of backup is recommended which could be stored on another computer automatically.
[Keywords] DSpace Institutional Repository Backup Recover

1 引言

用于构建机构仓储的DSpace系统软件，由于其开源，易安装，易维护，功能强大等特点，被国际上众多机构所采用，在DSpace Wiki上已经登记的机构就已经达到了 218 家[1]。目前，国内越来越多的高校和有关机构也正在运用DSpace软件构建本机构的仓储系统[1]。

当基于 DSpace 软件构建的机构仓储系统安装完毕并正式运行之后，随着数据量的剧增，仓储系统相应的维护管理工作量也开始增加。其中就包括了需要及时对系统进行备份，以便在仓储系统出现错误或崩溃时，能够利用备份文件准确、快速地进行恢复。

2 备份原因分析

由于各种不可避免或不可预知的原因，导致仓储系统数据丢失或运行出错，这时都需要通过预先的备份文件来进行及时恢复。这些原因大体可分为以下三类：

2.1 人为误操作

系统管理员或系统一般用户在使用过程中，由于人为误操作的原因，导致仓储系统失败或错误。比如在进行条目维护时误删除了条目数据；对 DSpace 软件系统进行改造时误删除了数据库中某个表；修改了表结构后系统出现异常；用户提交了大量条目到错误合集下，但

又不方便删除时，等等。

2.2 硬件系统失败

由于自然灾害、突然断电，及其他各种原因导致服务器系统的硬盘、内存、主板等物理硬件损坏，致使仓储系统数据丢失或异常。

2.3 其他原因

仓储系统在不同操作平台之间进行迁移时，比如从 Windows 操作系统迁移到 Linux 操作系统时，或者从测试平台迁移到正常发布平台时，都需要对仓储系统进行备份；对 DSpace 软件进行升级或改造维护之前，为确保数据安全，需要对仓储系统进行备份；为防范网络黑客的攻击与网络病毒入侵造成的损害，需要对仓储系统进行备份；等等。

3 制定备份策略

3.1 备份内容概述

从逻辑存储结构上看，存储在DSpace系统中的数据大约有 6 种类型，它们分别是：上传提交数据、用户信息、Dublin Core元数据与比特流格式注册信息、授权认可信息、系统使用日志信息和DSpace系统应用程序文件。其中最重要的上传提交数据的组织形式是社区（community）、合集（collection）、条目（item）、数字包（bundle）及比特流（bitstream）。其关系是比特流构成数字包，数字包构成条目，条目构成合集，合集构成社区[2]。相对于其他数据，提交到仓储系统中的数据的特点是数据量大、增长变化快，同时也是仓储系统最重要的数据。因此，上传提交数据是系统数据备份的重点。

从物理存储结构上看，DSpace 系统中的数据分别存储于三个目录文件中，它们分别是/dspace/（默认安装）、 [database]/（数据库数据目录）和[tomcat]/webapps/。上述的比特流数据、系统使用日志信息和部分 DSpace 系统应用程序文件存储在/dspace/目录下，大部分 DSpace 系统应用程序文件存储在[tomcat]/webapps/目录下，其余数据信息存储在数据库中。其中/dspace/目录下的数据文件和数据库中的数据信息对于仓储系统至关重要，是系统数据备份的重点。

3.2 备份策略

根据基于 DSpace 软件构建的机构仓储系统的数据结构类型、数据增长变化频率、数据重要性的轻重缓紧等情况，可以进行侧重于上传提交数据的备份，或者兼顾多类型数据的备份，或者兼顾系统数据与应用程序的备份。

4 备份与恢复的类型与方法[3]

4.1 使用导入与导出工具进行备份与恢复

DSpace 软件系统提供了数据的导入与导出脚本工具。用户可以使用该工具把系统中的数据导出或导入系统，实现仓储系统的备份与恢复。

4.1.1 适用环境

使用导入与导出工具进行系统数据备份与恢复方法，只用来备份与恢复上传提交数据，即条目及条目以下的数据信息。当把试验系统上的数据迁移到正式系统上时，可以采用此方法。

4.1.2 系统备份步骤

（1）确定需要备份的合集，并为每个合集创建一个存放备份文件的目录。

（2）使用导出工具，分别导出各个合集下的所有条目，导出工具命令如下[4]：

dsrun org.dspace.app.itemexport.ItemExport --type=COLLECTION/ITEM --id=collID/itemID --dest=dest_dir --number=seq_num

其中各参数含义如下：

Type—指定将要导出的是合集下的所有条目还是仅指单个条目，在此应指定为COLLECTION。

Id－指定将要导出合集或者条目的数据库 ID 或者 Handle。

Dest－指定导出的合集或者条目的存放目录。

Number－指定一个数字起始系列号。导出的每一条条目将自成一个文件夹，有若干个条目就有若干个文件夹。系列号用于文件夹的命名，并在起始系列号上递增。

4.1.3 系统恢复

（1）确定导入条目与合集的对应关系。

（2）使用导入工具，分别导入之前导出的条目到相应的合集中，导入工具命令如下[4]：

dsrun org.dspace.app.itemimport.ItemImport --add --eperson=joe@user.com --collection=collectionID --source=items_dir --mapfile=mapfile

其中各参数含义如下：

Add－指定将要导入条目到系统中。

Eperson－指定具有添加条目权限的用户 E-mail。

Collection－指定将要导入合集所在的数据库 ID 或者 Handle。

Source－指定将要导入条目的所在目录。

Mapfile－指定映像文件。

4.1.3 优劣势及注意点

（1）独立于数据库的备份与恢复，操作步骤简单，避免了有关数据库的备份与恢复的复杂操作。

（2）操作时不需要关闭系统，对系统运行不产生影响。但最好选择较少人提交数据的时段进行操作。

（3）当某个合集出现错误数据丢失时，如上传条目到了错误的合集中，但又不想重新上传时，可以采用此导入导出工具。

（4）由于导出的数据是一组格式规范的条目数据，一方面可以使用 DSpace 提供的导入工具把数据导入到原有系统中，另一方面，用户可以单独开发一种可以识别该格式的脚本程序，可以把条目导入到任何版本的系统中，即可以不依赖于原系统的结构而永久备份保存。

（5）使用导出与导入工具方法进行仓储系统备份与恢复，由于需要对每个合集逐个进行备份或恢复，实际操作起来比较繁琐。

（6）不能备份与恢复条目信息以外的相关信息，例如用户信息、Dublin Core 元数据与比特流格式注册信息、授权认可信息、系统使用日志信息和 DSpace 系统应用程序文件等。

（7）执行恢复操作时，操作者必须具有系统管理员权限或对应合集的管理员权限。

4.2 数据存储级备份与恢复

4.2.1 适用环境

备份数据库上的数据和比特流。在数据库中存储了仓储系统的用户信息、Dublin Core 元数据与比特流格式注册信息、授权认可信息等重要数据。仓储系统的比特流则默认存储在[dspace]/assetstore 目录下。

基于数据存储级的备份方法，可以最大程度上备份仓储系统中的数据。

4.2.2 系统备份步骤

（1）备份数据库中的数据

PostgreSQL数据库的备份命令如下[5]：

pg_dump dspacedb > dspacebak.dmp

其中各参数含义如下：

dspacedb－机构仓储系统的数据库名称。

dspacebak.dmp－备份文件名称。

Oracle数据库的备份命令如下[6]：

Exp system/password@dspace FULL=y FILE=dspacebak.dmp

其中各参数含义如下：

system－以 system 用户执行数据库导出备份操作。或者以具有 Exp_full_database 角色或 DBA 角色的用户执行此操作。

password－用户密码。

dspace－Oracle 数据库的 SID。

FULL－指定数据库导出模式，y 表示将导出除 sys 外所有其他方案的所有对象。

FILE－指定备份文件名称。

（2）备份 DSpace 系统比特流

比特流默认存储在/dspace/assetstore 目录中，为方便存储，用压缩工具对该目录进行压缩打包处理，命令如下：

```
tar -zcf assetstore.tar.gz /dspace/assetstore
```

（3）把生成的备份文件存储到本地专门的备份目录中或进行异地备份。

4.2.3 系统恢复

（1）把备份文件从备份目录中拷贝到相应的目录下。

（2）恢复 DSpace 系统比特流。用解压缩工具对压缩文件进行解压，命令如下：

```
tar -zxvf assetstore.tar.gz
```

（3）恢复数据库

PostgreSQL数据库的备份命令如下[5]：

psql dspacedb < dspacebak.dmp

Oracle数据库的备份命令如下[6]：

imp system/password@dspace FULL=y FILE=dspacebak.dmp

4.2.3 优劣势及注意点

（1）该方法可以较全面地备份仓储系统中的数据。

（2）该方法侧重于系统数据的备份。当数据库发生故障或者数据库需要重新安装或升级时，该方法可以快速恢复数据库的数据。

（3）当进行恢复时，必须确保现在的 DSpace 软件版本与数据备份之时的 DSpace 版本相一致。因为不同 DSpace 版本之间的数据库结构可能不一致。

4. 3 文件系统级备份与恢复

4.3.1 适用环境

文件系统级备份与恢复是指在操作系统层面上对仓储系统进行备份与恢复，而不涉及构成仓储系统的应用软件与数据库内容。对数据库系统的备份与恢复不熟悉，并且数据量的增长或者变化不是太快或者不是太重要的场合，以及希望简单操作的场合，都可以采用操作系统命令对 DSpace 应用软件及比特流数据，以及数据库进行备份与恢复。

4.3.2 系统备份步骤

（1）关闭 Tomcat 服务器

Sh [tomcat]/bin/shutdown.sh

（2）关闭数据库

关闭 PostgreSQL 数据库

pg_ctl -D [postgresql]/bin/postgres stop

.关闭 Oracle 数据库

SQL> shutdown immediate

（3）备份 DSpace 系统存储的比特流、配置文件、可执行文件、日志文件等等。默认安装情况下，此类文件都存储在/dspace 目录下。为方便存储，用压缩工具对该目录进行压缩打包处理。

tar -zcf dspace.tar.gz /dspace

（4）备份数据库数据文件

对于 PostgreSQL 数据库

tar –zcf databak.tar.gz [postgresql]/data

.对于 Oracle 数据库

tar -zcf oradatabak.tar.gz [oracle]/oradata

（5）打开数据库

对于 PostgreSQL 数据库

postmaster -D [postgresql]/data

.对于 Oracle 数据库

SQL> startup

（6）启动 Tomcat 服务器

Sh [tomcat]/bin/startup.sh

（7）把生成的备份文件存储到本地专门的备份目录中或进行异地备份。

4.3.3 系统备份步骤

（1）把备份文件从备份目录中拷贝到相应的目录下。

（2）用解压缩工具分别解压各个备份文件，并覆盖原来出错的文件，命令如下：

tar -zxvf dspace.tar.gz

tar –zxvf databak.tar.gz

tar -zxvf oradatabak.tar.gz

4.3.4 优劣势及注意点

（1）该备份方法相当于数据库的“冷备份”，数据备份全面，操作简单方便，恢复工作快捷。

（2）执行文件系统级备份与恢复时，需要关闭系统，可能对使用中的用户有影响。

（3）进行系统恢复时，现有系统平台必须与备份之时的系统平台相一致。

（4）执行上述各操作命令时，必须切换成有相应权限的用户进行操作，最好是 Root 用户。

4. 4 一种简单实用的系统备份方法

无论采用上述各种备份与恢复方法的哪一种方法，在实际应用过程中都各有其优缺点。仓储系统在实际的运行过程中，其数据量的增加或者改变，除非在早期运用过程中会比较大量外，平时提交到仓储系统中的数据量并不是太大，对仓储系统中的数据改动更是小量，而且这些变化绝大部分是在白天完成的。因此针对这各情况，可以在每天的凌晨时分对仓储系统进行“文件系统级备份”。因为在这个时段，一般没有数据操作，可以默认为仓储系统已经关闭，从而实现文件系统级备份。

下面给出能够自动进行备份，并把备份文件上传到 FTP 服务器，实现异地备份的简单

实用的文件系统级备份方法。

前提假设：操作系统为 Linux；DSpace 软件为默认安装方式；Tomcat 和数据库（PostgreSQL）安装在/usr/local/目录下。由于备份操作涉及多个用户权限转换问题，在此以 root 用户进行操作。

（1）编写备份命令脚本

在 root 用户默认目录下，新建文件 autobak，用 vi 编辑该文件，插入如下代码：

```
cd /
tar -zcf dspace`date +%w`.tar.gz dspace
cd /usr/local
tar –zcf pgsql`date +%w`.tar.gz pgsql
tar –zcf tomcat`date +%w`.tar.gz tomcat
cd /root
./autoftp
cd /
rm dspace`date +%w`.tar.gz
cd /usr/local
rm pgsql`date +%w`.tar.gz
rm tomcat`date +%w`.tar.gz
```

保存该文件后，使用命令 chmod u+x autobak，使该文件成为可执行文件。

（2）编写文件自动 FTP 上传命令脚本

在 root 用户默认目录下，新建文件 autoftp，用 vi 编辑该文件，插入如下代码：

```
#!/bin/sh
WW = `date +%w`
pre_dspace = "dspace"
pre_ pgsql = " pgsql "
pre_tomcat = "tomcat"
dspace = "${pre_dspace}${WW}.tar.gz"
pgsql = "${pre_ pgsql}${WW}.tar.gz"
tomcat = "${pre_tomcat}${WW}.tar.gz"
echo "open xxx. xxx. xxx. xxx
user username password
binary
hash
lcd /
put ${dspace}
lcd /usr/local
put ${pgsql}
put ${tomcat}
bye
" | ftp -n
```

保存该文件后，执行命令 chmod u+x autobak，使该文件成为可执行文件。Autoftp 文件可以把打包后的备份文件自动通过 FTP 上传到指定的 FTP 服务器上，同时自动把备份文件名改名为带有当日星期数的名称。例如 dspace3.tar.gz，其中的“3”即为星期三的意思，表示该文件在星期三生成。

（3）配置备份命令脚本自动执行

Linux 系统的 cron 进程可以在指定的时间点执行计划任务。利用这一特点，可以在计划任务中加入自动执行备份任务，使用 crontab –e 命令编辑计划任务文件，插入如下代码：

Auto backup DSpace , Tomcat, and PostgreSQL at 2:00 every day

0 2 * * * /root/autobak

通过上述设置，备份命令脚本将在每天凌晨 2 点自动被执行。

上述自动备份方法做好后，不需要人工干预，实现自动异地备份。自动备份方法每天生成一个备份，每个备份可以保存 7 天。

5 结语

基于 DSpace 软件构建的机构仓储系统，其最大的数据特点是系统重要数据分为数据库中数据与数据库外数据，两种数据必须同步。这样的特点使得仓储系统不能仅单纯靠数据库的方法来实现系统的备份与恢复。也是这种特点使得无论使用前述的哪种备份与恢复方法，在真正执行恢复时，可能会出现数据丢失的情况。为确保损失减少到最少程序，只能勤加备份。

参考文献

1 DspaceInstances.http://wiki.dspace.org/index.php/DspaceInstances（Accessd Apr.28,2007）

2 陈和.DSpace 系统与厦门大学机构存储的构建.数字图书馆论坛,2006,(9):61-67,75

3 BackupRestore.http://wiki.dspace.org/index.php/BackupRestore（Accessd Apr.28,2007）

4 Test Transform Docs from LaTex.http://wiki.dspace.org/index.php/Test_Transform_Docs_from_LaTex#Item_Importer_and_Exporter（Accessd Apr.28,2007）

5 PostgreSQL 7.4: Backup and Restore.http://www.postgresql.org/docs/7.4/static/backup.html（Accessd Apr.28,2007）

6 王海亮，王海凤等，精通 Oracle 10g 备份与恢复。北京：中国水利水电出版社，2005